CIND 820
Capstone Project
Literature Review

Name: Ibrahim Sayed Ahmed

Student Number: 501120722

Supervisor: Dr/ Uzair Ahmad

Date of submission: 31/10/2022

Abstract

Introduction:

Stroke, also known as brain attack, happens when blood flow to the brain is blocked, preventing it from getting oxygen and nutrients from it and causing the death of brain cells within minutes. According to the World Health Organization (WHO), it is the second cause of death worldwide after ischemic heart disease. Stroke victims can experience paralysis, impaired speed, or loss of vision. While some of the Stroke risk factors cannot be modified, such as family history of cerebrovascular diseases, age, gender and race, others can and are estimated to account for 60% - 80% of stroke risk in the general population. Therefore, predicting stroke outcome for new cases can be determining for them to be treated early enough and avoid disabling and mortal consequences. The purpose of this research is to create an accurate model for predicting Stroke outcome with Data Science and Machine Learning (ML) based on previous data and individual characteristics. Providing useful information for the medical staff to deploy the needed treatment and decrease risks and consequences.

Research Questions and Methods:

Research Question:

- What are the attributes related to stroke?
- What is the correlation between pre-existing health conditions and stroke risks?
- How to predict stroke risk?

The plan is to use data collected to identify the most important factors for stroke prediction.

The machine learning methods, techniques to be applied to address these questions are as follows:

- **What are the attributes related to stroke?**

  Using Exploratory Data Analysis

- **Implementation of supervised ML classification Algorithms**

  Building several models and compare which will be the best, Like Decision Tree Classifier (DT), Random Forest Classifier (RF), K-Nearest Neighbour (KNN)

Theme:

Classification and Predictive Analytics.

Objective:

The purpose of this project is to apply Classification and Predictive Analysis on the data:

- Identify the most important factors for stroke prediction.
- Visualize the results
- Propose a predictive analytics approach for stroke prediction.

LITERATURE SURVEY

This study looked at how machine learning techniques were used on stroke patients to predict stroke, which will also help to understand and resolve the problem in more effective ways.

This section analyses and reviews the previously published papers that deal with the work on stroke disease prediction using different machine learning approaches and algorithms.

José Alberto [REF 1] proposed Stroke prediction through Data Science and Machine Learning Algorithms. He used classification algorithms-decision Tree, Naive Bayes, Logistic regression, random forest, SVM, Neutral Network and others for predicting the stroke with related attributes. He compared these techniques and chose Random Forest to be the best performing ML algorithm for this specific problem. As for many models for disease prediction, the cost of having false negatives is much higher than the one if predicting false positives. This is because, even though the person is not prone to suffering from a stroke, taking precaution and the recommended treatment has low negative impact in persons health. On the other hand, if we suppose that we say that the person has not risk of having such disease, we can even put its life in danger. Therefore, it is very important for the model to be evaluated depending on the Business problem to be solved. The model resulting from this research shown a very low rate of False negative predictions, so it seems that this model works with good performance.

Jeena et al. [Ref 2] provides a study of various risk factors to understand the probability of stroke. It used a regression-based approach to identify the relation between a factor and its corresponding impact on stroke. She used Support Vector Machine (SVM). In this work, we have implemented SVM with different kernel functions and found that linear kernel gave an accuracy of 90 %.

A predictive analytics approach for stroke prediction using machine learning and neural networks research [Ref 3] show using machine learning and neural networks in the proposed approach. They performed feature correlation analysis and a step wise analysis for choosing an optimum set of features. Three machine learning algorithms were implemented on a set of different features and principal components configurations. They found that neural network works the best. The accuracy and miss rate for this combination are 78% and 19% respectively.

Sudha. A [Ref 4] under the guidance of her professors N. Jaisankar & P Gayathra, proposed a stroke predictive Model using classification techniques. They used classification algorithms-decision Tree, Naive Bayes & Neural Networks for predicting the stroke with related attributes. They utilized principal component analysis algorithm for dimension reduction. They studied & used sensitivity 7 accuracy indicators for evaluation. Decision tree achieved 95.29% of sensitivity & 98.01% of accuracy. Bayesian classifier achieved 8710% & 91.30% respectively. They compared these techniques and chose the decision tree as the best classification Method.

Department of Computer Architecture [Ref 5] and Automation team of Universidad Completeness de Madrid, Spain along with hospital Universitario de La Princesa, Madrid, Spain researched about the testing the hypothesis that state of art machine learning based modeling methods. They investigated if non-invasive monitoring technologies could aid in stroke type diagnosis. These methods can even be used to predict risks in the future, such as the patient's ultimate mortality. With 7 predictors and 2 goal variables—prediction of stroke type and prediction of death—they gathered a dataset made up of the medical records of 119 patients. They assessed over 6 distinct metrics using 7 different machine learning algorithms, including Decision Tree, KNN, Logistic Regression, Naive Bayes, Neural Network, Random Forest, and Support Vector Machines. Additionally, they used the 10-fold cross validation re-sampling technique for the trained classifier's validation against any unknown sample and a guaranteed validation set from training one. Sensitivity, specificity, accuracy, the F measure, and areas under ROC as well as PRC were the model performance evaluation metrics utilized to compare the algorithms. With values of 0.93+0.03 and 0.970.01, respectively, Random Forest models produced the greatest results overall in the diagnosis of stroke prediction and death prediction when compared to other algorithms.

A team of National Institute of Engineering, Karnataka, conducted a survey about AI applications in stroke and aimed to predict the accurate results of occurrence of stroke. They used predictive algorithms and parameters that include patients' characteristics like gender, age, height, BMW, etc., they built a data model using decision tree algorithm to analyze these parameters. The result was analyzed using confusion matrix and the accuracy was 95%. To achieve this, they built the training model that helped to compare the newly fed data with the survey data. And the report was generated based on this comparison.

The outcomes of the various techniques show that a variety of factors can have an impact on the results of any study that has been performed. Any study's conclusion will be influenced by these numerous aspects, which include the method of data collection, the features used, the method used to clean the data, the method used to impute missing values, randomization, and standardization of the data. It is crucial that the researchers understand the relationships between the different input factors in an electronic health record and how they affect the final stroke prediction accuracy.

Data Description and EDA (Exploratory Data Analysis)

The dataset to be used for this project are Stroke Prediction Dataset.

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

The dataset has about 5k records. There are 12 columns: -id - gender - age - hypertension - heart_disease - ever married - work_type - Residence_type - avg_glucose_level - bmi - smoking_status – stroke.

Attribute Information

1) id: unique identifier

2) gender: "Male", "Female" or "Other"

3) age: age of the patient

4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

6) ever_married: "No" or "Yes"

7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"

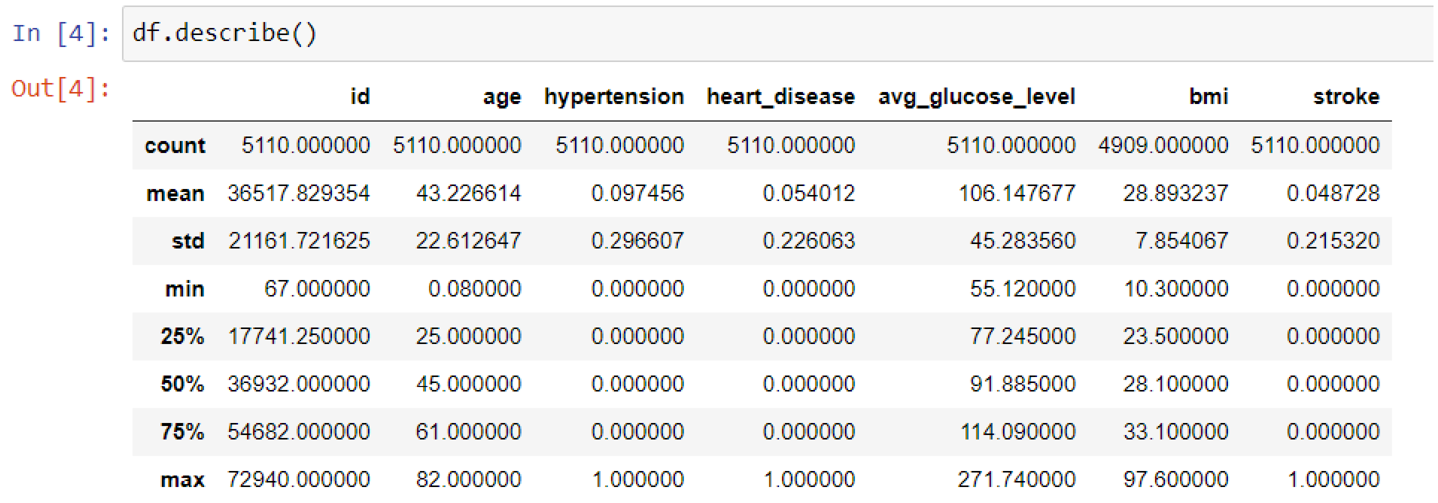
8) Residence_type: "Rural" or "Urban"

9) avg_glucose_level: average glucose level in blood

10) bmi: body mass index

11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*

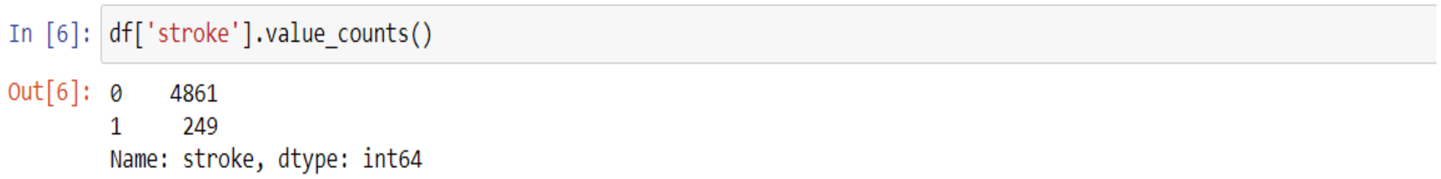
12) stroke: 1 if the patient had a stroke or 0 if not



```
In [4]: df.describe()
```

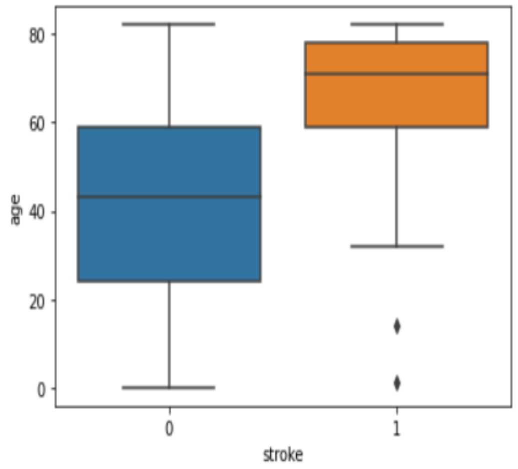
Out[4]:

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5110.000000	5110.000000	5110.000000	5110.000000	5110.000000	4909.000000	5110.000000
mean	36517.829354	43.226614	0.097456	0.054012	106.147677	28.893237	0.048728
std	21161.721625	22.612647	0.296607	0.226063	45.283560	7.854067	0.215320
min	67.000000	0.080000	0.000000	0.000000	55.120000	10.300000	0.000000
25%	17741.250000	25.000000	0.000000	0.000000	77.245000	23.500000	0.000000
50%	36932.000000	45.000000	0.000000	0.000000	91.885000	28.100000	0.000000
75%	54682.000000	61.000000	0.000000	0.000000	114.090000	33.100000	0.000000
max	72940.000000	82.000000	1.000000	1.000000	271.740000	97.600000	1.000000



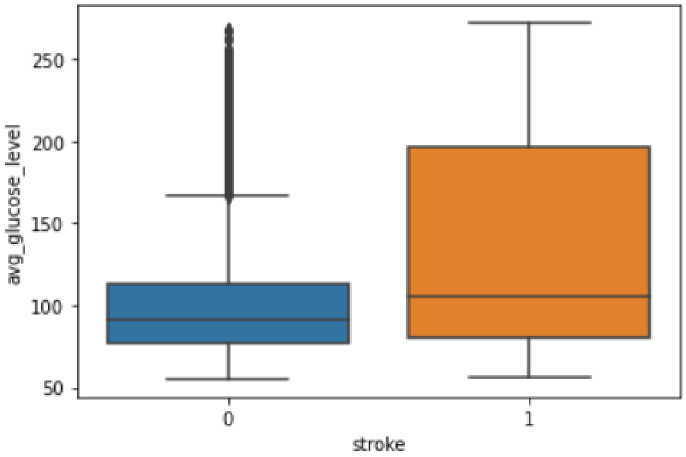
```
In [6]: df['stroke'].value_counts()

Out[6]: 0    4861
        1     249
        Name: stroke, dtype: int64
```

this dataset was highly unbalanced. Only 249 patients suffered a stroke while the remaining 4861 patients did not have the experience.

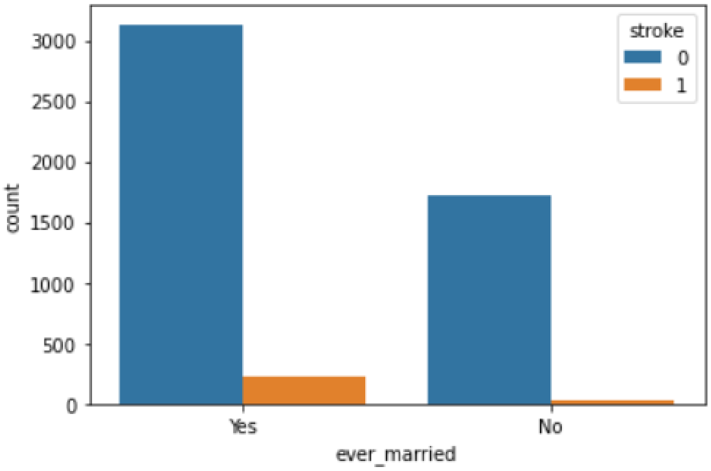


Strokes are more likely to occur in people over 60.

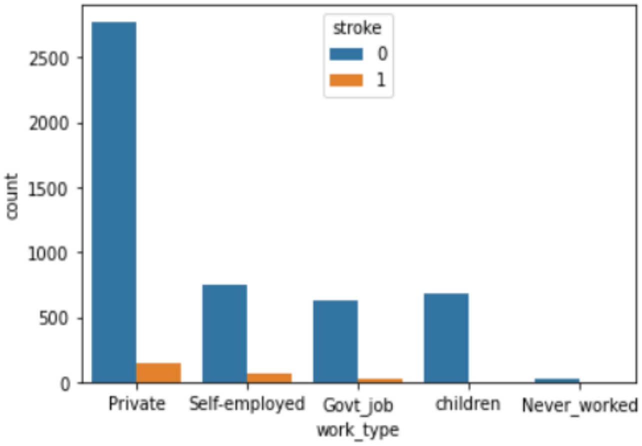
Strokes are more likely to occur in people over 60. Some anomalies can be identified as strokes occurring in people under the age of 20. Given that our food and living habits have an impact on stroke, it's probable that the data is accurate. Another finding is that those over 60 years old make up the group of people who do not experience strokes.



people having stroke have an average glucose level of more than 100.



People who are married have a higher stroke rate.

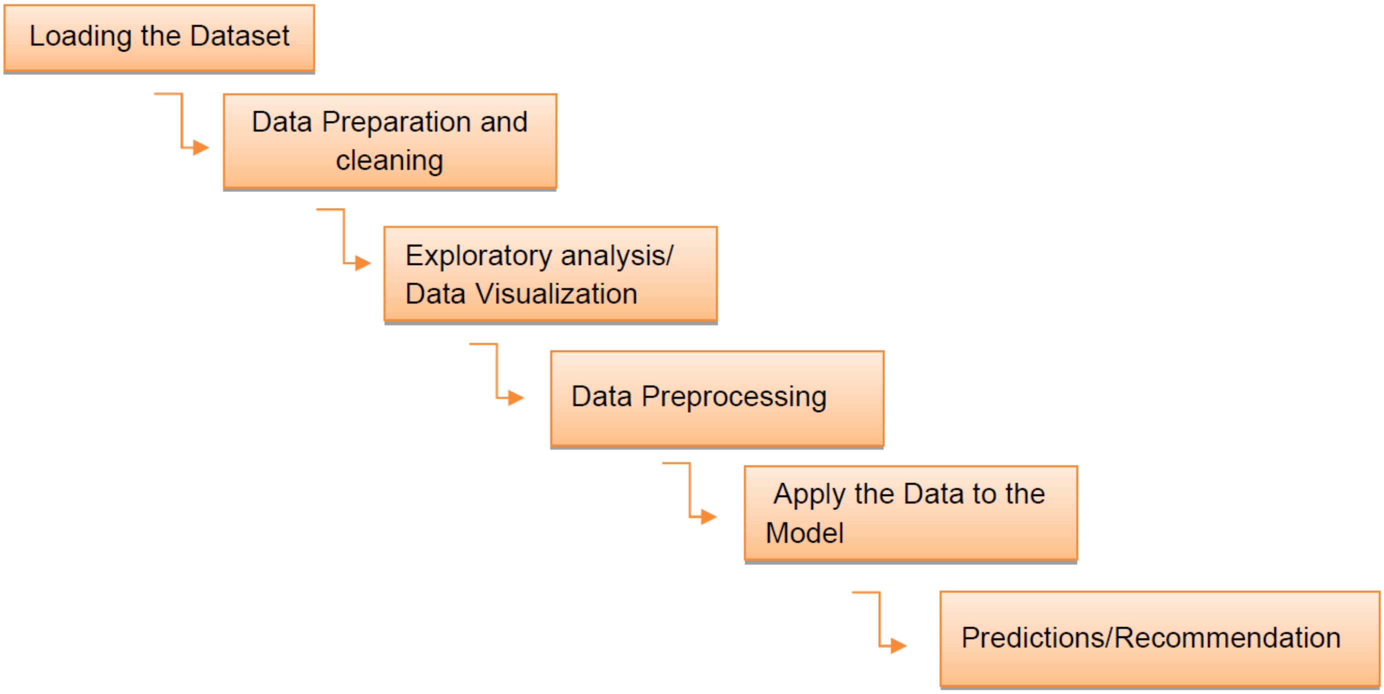


People working in the Private sector have a higher risk of getting a stroke. And people who have never worked have a very less stroke rate.

CONCLUSION

The ability to predict stroke in a variety of situations has benefited from numerous machine learning techniques. Considering scenarios, data sets, parameters, and other analysis should be the basis for every decision about the usage of a certain machine learning technique. On the best technique to employ for stroke prediction, we cannot come to any conclusions. There are benefits and drawbacks to each strategy. Depending on the importance of the specific problem statement, it is wise to select one of them. To choose the precise technique or model to apply, one needs undertake statistical analysis and initialization. Random forest, however, is one of the most well-liked and effective methods for predicting a quantity from a sample of data since it produces promising outcomes.

Methodology



Loading the Dataset
Data Preparation and cleaning
Exploratory analysis/ Data Visualization
Data Preprocessing
Apply the Data to the Model
Predictions/Recommendation

A link to repository on GitHub

https://github.com/Ibrahim-S-Ahmed/Stroke_Prediction_Capston_Project

References

1. https://www.researchgate.net/publication/352261064_Stroke_prediction_through_Data_Science_and_Machine_Learning_Algorithms
2. Stroke prediction using SVM | IEEE Conference Publication | IEEE Xplore
3. https://www.sciencedirect.com/science/article/pii/S2772442522000090
4. https://www.semanticscholar.org/paper/Effective-Analysis-and-Predictive-Model-of-Stroke-Sudha-Gayathri/d46f2f63cc3ca9714c767b805cd033776f0bd3ee

5. https://www.researchgate.net/publication/335515560_Comparison_of_Different_Machine_Learning_Approaches_to_Model_Stroke_Subtype_Classification_and_Risk_Prediction
6. https://www.hindawi.com/journals/bn/2022/7725597/
7. https://www.ijraset.com/research-paper/automated-prediction-of-brain-stroke-disease-classification
8. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0234722